Core expansion in diachronic prototype semantics: a computational case study on loanwords

Stefano De Pascale
KU Leuven & Vrije Universiteit Brussel, stefano.depascale@kuleuven.be

Keywords: vector space model, semantic change, Dutch, prototype semantics

In recent years, the computational modeling of semantic change has witnessed an enormous growth, testified by the introduction of new techniques (Kutuzov et al. 2018). Deep neural networks called 'transformers' have become the state of the art in the computational modelling of semantics in contemporary texts and are now also applied to historical corpora (Manjavacas & Fonteyn 2022). The gist of these models is that they represent the meaning of a single corpus occurrence (i.e. a token) as a single numerical vector and vectors of many tokens are then compared through similarity indices in order to assess whether they share the same meaning or not. Hence the term token-based vector representations.

There is also a growing awareness that the application of token-based vector representations on historical data needs grounding in historical semantic scholarship, and that consequently a tighter connection between computational modeling and linguistic theorizing should be pursued (Tahmasebi et al. 2021). Among the cognitive-functional realm of theories, diachronic prototype theory (Geeraerts, 1997) is the most appropriate framework to focus on and to put to test. At the same time it is most amenable to be tested by means of token-based vector representations based on corpora, as its key tenets are all very well suited to be subjected to computational operationalization (e.g.: the importance of quantitative aspects of semantic structure; the blurring of the distinction between semantic and encyclopedic knowledge etc.)

In this presentation I offer a case study on a fundamental hypothesis in the diachronic-prototype semantic literature, namely that changes in the referential range of one specific word meaning may take the form of a peripheral expansion on the prototypical core cases within that referential range (Geeraerts 1997: 23). This hypothesis relies on the distinction between the prototypical core and the periphery of a conceptual category. Crucially, this core/periphery distinction can be formulated in terms of token-based vector representations: core applications are those that will have the most tokens in a certain time frame and those tokens will contain the highest number of similar context words; peripheral applications will have less tokens which contain less similar context words relative to the core.

Loanwords are a suitable test case for this expectation, as they usually enter the language with a simple, monosemous reading, so that allows us to keep track in detail of the development of a word. As test cases I take a selection of 9 English loanwords that entered into the Dutch language during the 20th century, and as a corpus I rely on the newspaper section of the historical Delpher corpus (archived by the National Library of the Netherlands). Concretely, I focus on the frequency changes and vector-similarity changes of the context words of the tokens of those loanwords. I expect to see that the shift to major new meanings (either core senses or secondary, figurative senses) will usually take place after an initial expansion of the original core meaning. If core cases are indeed more stable, I expect to see earlier movement in the center than in the periphery.

References

Geeraerts, Dirk. 1997. *Diachronic prototype semantics*. Oxford: Oxford University Press.

Kutuzov, Andrey, Øvredlid, Lilja, Szymanski Terrence, & Velldal, Erik. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1384–1397. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Manjavacas Arevalo, Enrique & Fonteyn Lauren. 2022. Non-Parametric Word Sense Disambiguation for Historical Languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, 123–134. Taipei, Taiwan: Association for Computational Linguistics.

Tahmasebi Nina, Borin Lars, & Jatowt Adam. 2021. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 1–91. Berlin: Language Science Press.